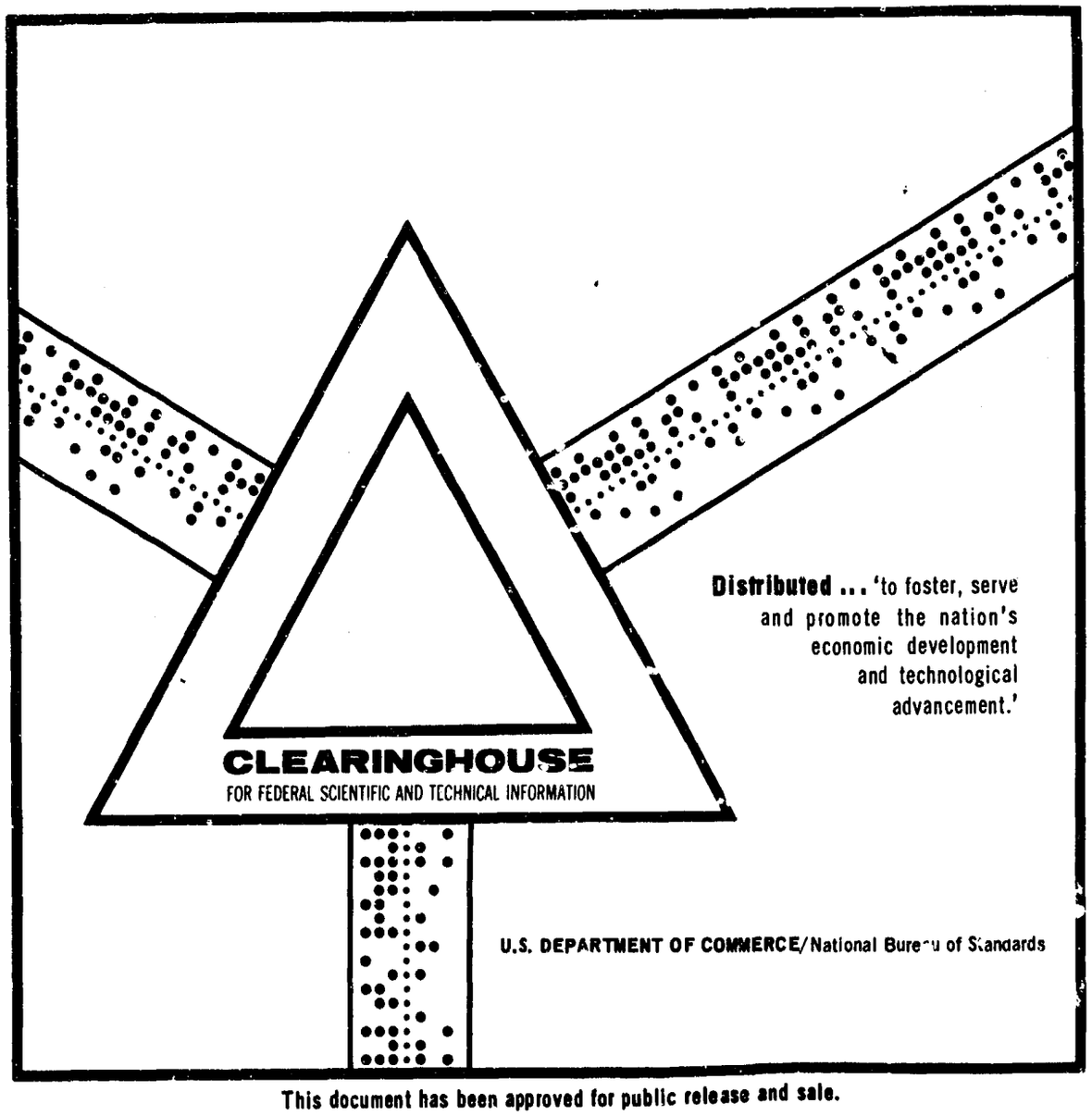
RANDOMIZATION OF SYMBOL REPETITION OF PUNCH CARDS WITH SUPERIMPOSED CODING IN INFORMATION-SEARCH SYSTEMS

L. Ya. Pirovich

Foreign Technology Division
Wright-Patterson Air Force Base, Ohio

24 March 1969



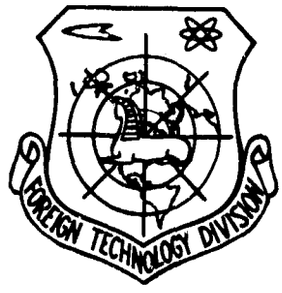
FTD-HT-23-881-68

FOREIGN TECHNOLOGY DIVISION

RANDOMIZATION OF SYMBOL REPETITION OF PUNCH CARDS WITH SUPERIMPOSED CODING IN INFORMATION-SEARCH SYSTEMS

by

L. Ya. Pirovich



EDITED TRANSLATION

RANDOMIZATION OF SYMBOL REPETITION OF PUNCH CARDS WITH SUPERIMPOSED CODING IN INFORMATION-SEARCH SYSTEMS

English pages: 10

Sources: Nauchno-tekhnicheskaya informatsiya. Seriya 2. Informatsionnyye protsessy i sistemy, No. 7, 1967, pp. 21-23.

Translated by: Contract No. F33657-68-D-1287

PREPARED BY:

TRANSLATION DIVISION
FOREIGN TECHNOLOGY DIVISION
WP-AFB, OHIO.

DATA HANDLING PAGE

01-ACCESSION NO. 98-DOCUMENT LOC TP9000391	39-TOPIC TAGS information storage and retrieval, punched card, coding evaluation, data processing system
09-TITLE RANDOMIZATION OF SYMBOL REPETITION OF PUNCH CARDS WITH SUPERIMPOSED CODING IN INFORMATION-SEARCH SYSTEMS	
47-SUBJECT AREA 05, 09	

42-AUTHOR/CO-AUTHORS PIROVICH, L. Ya.	10-DATE OF INFO 23JUL67
43-SOURCE NAUCHNO-TEKHNICHESKAYA INFORMATSIYA. SERIYA 2. INFORMATSIONNYYE PROTSESSY I SISTEMY (RUSSIAN)	68-DOCUMENT NO. FTD-HT-23-881-68
	69-PROJECT NO. 6050205

63-SECURITY AND DOWNGRADING INFORMATION UNCL, O	64-CONTROL MARKINGS NONE	97-HEADER CLASN UNCL

76-REEL/FRAME NO. 1888 1106	77-SUPERSEDES	78-CHANGES	40-GEOGRAPHICAL AREA UR	NO OF PAGES 10
CONTRACT NO. F33657-68-D-1287	X REF ACC. NO. 65-	PUBLISHING DATE 94-00	TYPE PRODUCT TRANSLATION	REVISION FREQ NONE
STEP NO. 02-UR/0447/67/000/007/0021/0023			ACCESSION NO.	

ABSTRACT

(U)The article shows the effect of the irregularity of using separate symbols on search noise on punch cards with superimposed symbol coding in information-search system (IPS). A binomial law of random value distribution of repetition of each symbol is established and analyzed. A method of determining the maximum value of symbol repetition is proposed and an example of calculating this value for an experimental IPS is given.

RANDOMIZATION OF SYMBOL REPETITION ON PUNCH CARDS WITH SUPERIMPOSED CODING IN INFORMATION-SEARCH SYSTEMS

L. Ya. Pirovich

The article shows the effect of the irregularity of using separate symbols on search noise on punch cards with superimposed symbol coding in information-search system (IPS). A binomial law of random value distribution of repetition of each symbol is established and analyzed. A method of determining the maximum value of symbol repetition is proposed and an example of calculating this value for an experimental IPS is given.

The use of superimposed symbol coding on cards with edge perforations and the use of slotted and machine punch cards when creating mechanized information-search systems (IPS) is connected with the appearance of superfluous punch cards, not responding to the search interrogation during search [1-3]. The superimposition of the codes is determined by the fact that several symbols are entered on a single code field of the punch card. The irregularity of encountering separate symbols in different documents has a substantial effect on search noise[1] of this system. Single symbols or groups of symbols are encountered more often than others both during indexing as well as during search.

We know [4, 5] that code configurations of often repeated symbols decrease the selectivity of the codes, containing code symbols common to them[2].

As an illustration, let us give an example from a paper [5]. The symbols in the IPS

strengthen the TVCh
hardness of the HRC 45-50
the diameter of the information slot is 40 to 50 mm

often fall together on a single punch card.

Let us write their code configurations:

---

[1] See p. 9.
[2] See p. 9.

65—44—[08]

72—[28]—14

[79]—36—13

The code symbols of these configurations form many new codes. One of them is separated by squares. Recovery of the punch cards according to this code configuration will, obviously, yield a very large number of superfluous cards.

The presence of often repeated symbols and their entry onto a single code field of the punch card leads to an increase of the average search noise of the system and, in isolated cases, yields an especially large number of superfluous punch cards.

In order to decrease search noise, it is necessary to randomize symbol repetition of the IPS, to exclude often repeated symbols from the total code field of a punch card, and to enter them on specially isolated fields.

In connection with this, documentation specialists working with the IPS will be interested in the acceptable level of symbol repetition, above which the search noise of the system will be increased and will exceed established limits. Paper [5] describes an IPS, which provides for the exclusion of symbols from the overall code field of the punch card according to the increase of their repetition rate. In this case, it is also very important to know the acceptable extent of symbol repetition, upon achievement of which the symbols will become more repetitious.

When setting up the main part of the punch cards of the IPS, the information specialist analyzes each document in the IPS storage and allots to it a group of characteristic symbols (descripters), available in the dictionary (the register) of the system. The total of symbols selected is the search image of the document and is entered on the punch card.

Thus, all documents entering the IPS storage are processed. The appearance of one or another symbol is a random event; therefore, we shall approach investigation of such events from the probability theory.

Let us assume that there are n symbols in the IPS dictionary, that an average of r symbols is entered onto each punch card and that the main part of the punch cards consist of Q cards.

Isolation in the document of the total of random symbols can then be presented, using standard models of probability theory, as the random selection of symbols of volume r from a general total of n symbols. Selection of a volume r is non-recurring selection, because only r different symbols can be entered onto the punch card and no symbol can be entered twice or more than twice[3]. Moreover, the order of symbol distribution in this selection makes no difference.

The number of such different selections from a general aggregate n is equal to the number of combinations of n symbols according to r (C_n^r).

Let us agree to consider equally possible the appearance of any of the random selections of a volume r with a probability equal to $1/C_n^r$.

Let us select at random one of the n symbols and investigate a principle of its repetition during compilation of the main part of Q cards.

The probability that the symbol being studied will fall in any of the selections of volume r, and, consequently, on any punch card is equal to

$$p = 1 - \frac{C_{n-1}^r}{C_n^r}.$$

Solving this equation, we obtain

$$p = \frac{r}{n}. \tag{1}$$

A series of Q independent Bernoulli tests are performed, each of which is a selection from the overall total n of random selection of volume r. The probability that a symbol will appear in the selection is equal to p, the probability that it will not appear is equal to q = 1 - p.

Let us determine the probability that the symbol being studied will appear K times in Q independent tests. Let us denote this probability by $p_{K,Q}$ and let us write [7]:

$$p_{K,Q} = \frac{Q!}{K!(Q-K)!} p^K q^{Q-K}. \tag{2}$$

Substituting the values p and q in equation (2), we obtain

$$p_{K,Q} = \frac{Q!}{K!(Q-K)!} \left(\frac{r}{n}\right)^K \left(1 - \frac{r}{n}\right)^{Q-K}. \tag{3}$$

---

[3] See p. 9.

The number K in the series of Q tests is random and may assume the following whole numerical values

$$0;\ 1;\ 2;\ \ldots;Q-1;\ Q$$

with the corresponding probabilities

$$P_{0,Q};\ P_{1,Q};\ P_{2,Q};\ \cdots;\ P_{Q-1,Q};\ P_{Q,Q}.$$

Equation (3) determines the binomial probability distribution of random value K.

According to [7] the mathematical expectancy (average value) and the average quadratic deviation (the degree of scattering of the values of K near the mathematical expectancy) for a random value of K comprises:

$$M(K)=Qp \tag{4}$$

$$\sigma(K)=\sqrt{Qpq}. \tag{5}$$

Substituting the values p and q in equations (4) and (5), we obtain

$$M(K)=\frac{Qr}{n} \tag{6}$$

$$\sigma(K)=\sqrt{\frac{Qr}{n}\left(1-\frac{r}{n}\right)}. \tag{7}$$

On the condition of the equal probability of random selections of volume r, we hypothetically established the binomial distribution law of the discrete random value of repetition of any arbitrarily selected symbol of the IPS, and we determined its mathematical expectancy and average quadratic deviation.

The IPS sketches of mechanical components were used for a statistical check of the hypothesis on the binomial distribution law of the random value of symbol repetition [5].

A sub-group of symbols "lengths of components or surfaces in mm," containing 45 symbols, was isolated from the IPS. It is almost equally possible that these symbols will be encountered on the documents.

Further, punch cards were selected on which symbols from this sub-group were found. The real value of repetition of each of the 45 symbols was computed.

As a result of statistical processing of the data obtained [8], a polygon of distribution of the discrete random value of symbol repetition was plotted (Fig. 1).

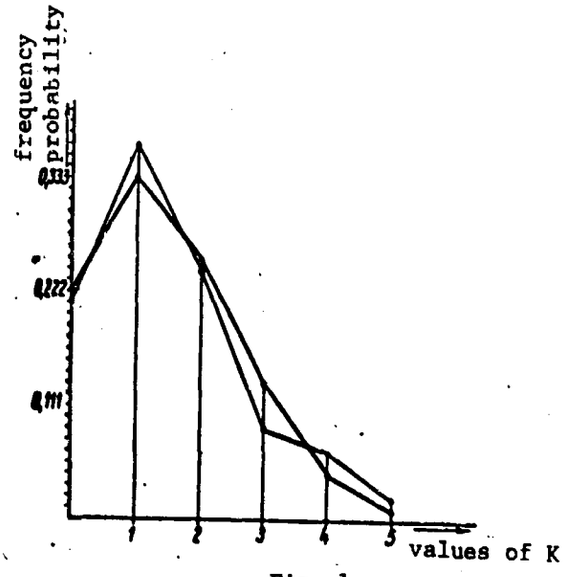


Fig. 1

The values of the symbol repetition value K were plotted on the axis of the abscissae, and corresponding empirical frequencies of the random value K, obtained in 90 experiments, and the theoretical probabilities, computed by formula (3), were plotted on the axis of the ordinates.

The thin broken line in Fig. 1 is the graph of the empirical frequency distribution of the random value of K. The thick broken line is the graph of the distribution of theoretical probabilities of the random value of K.

The probability that the empirical and theoretical graphs of distribution will coincide according to Pearson's agreement criteria X^2 at four orders of freedom is equal to 0.486. This probability is quite large; therefore, the hypothesis that the random value of symbol repetition is distributed according to a binomial law may be considered plausible.

According to the Moire-Laplace limiting theorem [9] according to the degree of increase of a number of independent Bernoulli tests, the binomial distribution asymptotically approaches the normal, having the same mathematical expectancy and average quadratic deviation.

In our case, the number of independent tests is equal to the number of

punch cards Q in the main part of the IPS. Usually, the main part of hand sorted punch cards includes from several hundred to several thousand cards. The main part of machine sorted punch cards may consist of several tens of thousands of cards (we recommend up to 100,000).

At a given probability p that the symbol being studied will be entered on the punch card, we can determine the minimum number of punch cards Q_{min} in the IPS, which approximately permits the application of normal distribution, rather than the binomial. In order to do this, we use the inequality given in paper [10]:

$$Q_{min} \geq \frac{9}{p(1-p)} . \qquad (8)$$

In the interval $0 < p < 0.1$, Q is almost inversely proportional to p.

Let us consider Q_{min} for an experimental IPS as an illustration of the given equation [5].

The IPS parameters are the following: n = 139 symbols, r = 6 symbols, according to formula (1)

$$p = \frac{6}{139} \simeq 0{,}043.$$

Hence,

$$Q_{min} > \frac{9}{0{,}043 \cdot 0{,}957} \simeq 220 \text{ card.}$$

Obviously, an approximate normal distribution of the random value of symbol repetition is suitable for most IPS.

This circumstance makes it possible to use the well-studied normal distribution law, which has been described in detail in the literature, for analysis of the extent of symbol repetition.

According to the normal distribution law [11], 99.73% of all values of the random value K, that is, practically all values of the repetition value of any symbol of the IPS, must be located in the range from $M(K) - 3\sigma(K)$ to $M(K) + 3\sigma(K)$. An increase in the number of repetition of separate symbols has an effect on the increase of search noise of the system. Therefore, we are interested in the upper acceptable limit of the extent of symbol repetition, at which the average search noise of the system will not exceed

that established for a given IPS level. The normal distribution law provides an answer to this problem and permits us to establish the quantitative character of the distribution value of symbol repetition.

Thus, 13.59% of the repetition value of each symbol, should be located in the range from $M(K) + \sigma(K)$ to $M(K) + 2\sigma(K)$, and only 2.14% -- in the range from $M(K) + 2\sigma(K)$ to $M(K) + 3\sigma(K)$ (Fig. 2).

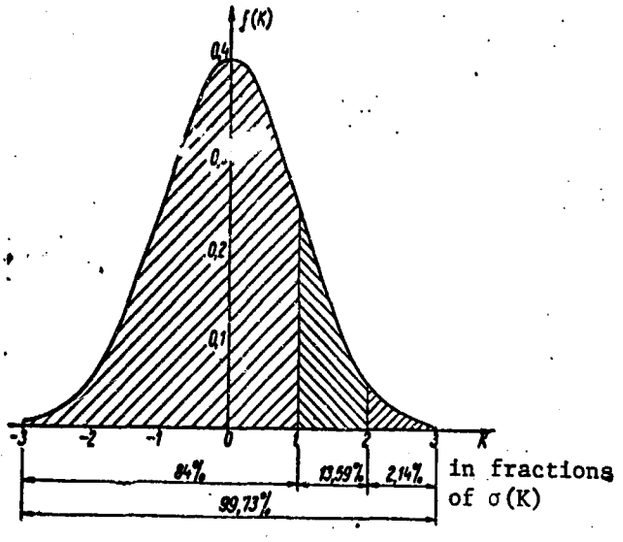


Fig. 2

Imagine that the repetition values K of all n equally-possible symbols are a frequency of the values of the random value K of one symbol in n experiments. Then, relying on the given positions of the normal distribution law, we compile a correlation table, in which will be shown the dependence between assumed greatest repetition values and the number of IPS symbols corresponding to them.

Correlation Table

Greatest value of symbol repetition, K_{max}	Number of symbols (in fractions of n)
from $M(K) + \sigma(K)$ to $M(K) + 2\sigma(K)$	0.136 (13.59%)
from $M(K) + 2\sigma(K) + 1$ to $M(K) + 3\sigma(K)$	0.021 (2.14%)

Let us cite an example from an experimental IPS [5].

The parameters of the IPS are as follows: Q = 177 punch cards, n = 139 symbols, and r = 6 symbols.

Using equations (6) and (7), we determine the mathematical expectancy and average quadratic deviation of the random value of repetition K of any symbol.

$$M(K)=\frac{177.6}{139}\simeq 8$$

$$\sigma(K)=\sqrt{\frac{177.6}{139}\left(1-\frac{6}{139}\right)}\simeq 3$$

Using the correlation table, we determine that 139.0.136 = 19 and 139.0.021 = 3 symbols should have maximum repetition 12 - 14 and 15 - 17 times, respectively.

Preserving in IPS [5] the specified proportions between the number of symbols and their repetition value, we guard the system against an increase of the established average search noise due to often repeated symbols[4].

Symbols and their random totals, encountered in documents are usually not equally probable in practice.

This is particularly true of often repeated symbols.

The proposed mathematical model of distribution laws of symbol repetition makes it possible to compute for any IPS according to its basic parameters: the number of punch cards (of documents) Q, the number of symbols n, and the average number of symbols on the punch card r, which is the maximum acceptable value of symbol repetition. Symbols which repeat the greatest number of times must be excluded from coding in the normal code field of a punch card.

Thus, the distribution of the values of symbol repetition of a real system artificially approaches the theoretical distribution.

---

[4] See p. 9.

Footnotes

1. To p. 1. Search noise [1] is the portion of those delivered, but unrelated to the given information interrogation of the punch cards (of the documents).

2. To. p. 1. The code symbol is the number of the position or cell of the punch card, from a random combination of which the code configuration (the code) of the symbol is formed.

3. To p. 3. Non-recurring selection according to paper [6] is selection in which a once selected element is separated from the overall total, because the selection contains no repeated elements.

4. To p. 8. The maximum symbol repetition obtained for the experimental IPS of [5] is somewhat inaccurate, i.e., instead of 220 punch cards, the minimum recommended for normal approximation of the binomial distribution is 177.

Bibliography

1. Mikhaylov, A. I., A. I. Chernyy and R. S. Gilyarevskiy, Osnovy Nauchnoy Informatsii [Fundamentals of Scientific Information], Nauka Press, Moscow, 1965.

2. Perforirovannyye Karty i ikh Primeneniye v Nauke i Tekhnike [Punch Cards and Their Application in Science and Technology], Mashgiz Press, Moscow, 1963.

3. Kent, A., Informatsionno-poiskovyye Sistemy [Information-search Systems], VNIIEM Press, Moscow, 1965.

4. Shul'ts, K. K., Primeneniye Sluchaynykh Kodov k Poisku Literatury (Application of Random Codes to Literature Search), Perforirovannyye Karty i Ikh Primeneniye v Nauke i Tekhnike [Punch Cards and Their Application in Science and Technology], Mashgiz Press, Moscow, 1963.

5. Pirovich, L. Ya., Informatsionno-poiskovaya Sistema Chertezhey Detaley (An Information-Search System for Design Components), NTI, No. 6, 1966.

6. Feller, V., Vvedeniye v Teoriyu Veroyatnostey i yeye prilozheniya [Introduction to Probability Theory and Its Application], Mir Press, Moscow, 1964.

7. Shor, Ya. B., Statisticheskiye Metody Analiza i Kontrolya Kachestva i Nadezhnosti [Statistical Methods of Analysis and Control of Quality and Reliability], Sovetskoye Radio Press, Moscow, 1962.

8. RTM44-62, Metodika Statisticheskoy Obrabotki Empiricheskikh Dannykh [A Method of Statistical Processing of Empirical Data], Publishing House of Standards, Moscow, 1966.

9. Gnedenko, B. V., Kurs Teorii Veroyatnostey [A Course in Probability Theory], Moscow, 1961.

10. Khal'd, A., Matematicheskaya Statistika s Tekhnicheskimi Prilozheniyami [Mathematical Statistics with Technological Applications], Foreign Literature Press, Moscow, 1956.

11. Smirnov, N. V. and I. V. Dunin-Barkovskiy, Kurs Teorii Veroyatnostey i Matematicheskoy Statistiki Dlya Tekhnicheskikh Prilozheniy [A Course in Probability Theory and Mathematical Statistics for Technological Applications], Nauka Press, Moscow, 1965.